A generalized likelihood ratio test to identify differentially expressed genes from microarray data

Song Wang* and Stewart Ethier

Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA

Received on April 3, 2003; revised on June 17, 2003; accepted on July 26, 2003

ABSTRACT

Motivation: Microarray technology emerges as a powerful tool in life science. One major application of microarray technology is to identify differentially expressed genes under various conditions. Currently, the statistical methods to analyze microarray data are generally unsatisfactory, mainly due to the lack of understanding of the distribution and error structure of microarray data.

Results: We develop a generalized likelihood ratio (GLR) test based on the two-component model proposed by Rocke and Durbin to identify differentially expressed genes from microarray data. Simulation studies show that the GLR test is more powerful than commonly used methods, like the fold-change method and the two-sample t-test. When applied to microarray data, the GLR test identifies more differentially expressed genes than the t-test, has a lower false discovery rate and shows more consistency over independently repeated experiments.

Availability: The approach is implemented in software called GLR, which is freely available for downloading at http://www.cc.utah.edu/~jw27c60

Contact: song.wang@m.cc.utah.edu

INTRODUCTION

Microarray technology, which can measure expression intensities of thousands of genes in a single experiment, becomes a powerful tool in molecular biology and medicine. Microarray technology can address many important questions in life science. One major application of microarray technology is to identify differentially expressed genes between samples. Detection of those genes can help scientists understand the mechanism of many diseases and find potential targets for new therapeutic strategies.

The great potential of microarray technology is currently limited by the lack of good data analysis tools (Nadon and Shoemaker, 2002). Analysis of microarray data is difficult for several reasons. First, the small number of samples relative to the large number of genes tends to cause low sensitivity and low specificity at the same time. Second, gene expression data depart rather dramatically from the assumptions required by traditional statistical methods, resulting in poor performance of these methods.

One straightforward approach widely used by biologists to detect difference in expression is the fold-change method. It considers genes whose expression intensities are different by more than 2-fold between control and experimental samples as significantly differentially expressed. Such a simple ‘two-fold rule’ is easy to calculate but seldom gives optimal results, because it does not take variation within treatments into consideration. This method tends to be too liberal when applied to lowly expressed genes and too conservative when applied to highly expressed genes.

Another method is the traditional two-sample t-test. The performance of the t-test depends on the sample size, and whether the expression intensities can be assumed as normally distributed. Since the level of replication within treatments is often low for microarray experiments and the expression intensities may not be normally or even symmetrically distributed, the performance of the t-test is usually poor.

Non-parametric methods like Mann–Whitney U-test and Wilcoxon rank test do not require distributional assumptions and are thus robust, but their powers are reduced, especially in the case of small sample sizes typical in microarray studies, they can detect far fewer genes than parametric methods like t-test (Thomas et al., 2001).

To develop a better statistical method to analyze microarray data, one needs better understanding of the distribution and the error structure of gene expression data. It is a long-known observation that the SD of the expression intensity of a gene increases with its mean. Chen modeled this relationship as a linear function, predicting a constant coefficient of variation throughout the whole range of expression levels (Chen et al., 1997). This model fits microarray data well only for genes with high expression levels, while for genes with low expression levels, the coefficient of variation is increasing and tends to infinity as the expression level approaches zero, suggesting the existence of a base line error that does

*To whom correspondence should be addressed.
not decrease with the mean. To describe better the error structure of gene expression data from microarray studies, Rocke and Durbin (2001) proposed a model that reflects two types of errors, one additive and one multiplicative. The model is

\[ y = \alpha + \mu e^{\eta} + \varepsilon \]  

(1)

where \( y \) is the measured expression intensity for a single gene, \( \alpha \) is the mean background, \( \mu \) is the true expression intensity, \( \eta \) represents the multiplicative error that is dominant at high-expression levels and \( \varepsilon \) represents the additive error that is noticeable mainly at low-expression levels. \( \eta \) and \( \varepsilon \) are assumed normally distributed with mean 0 and variance \( \sigma^2_{\eta} \) and \( \sigma^2_{\varepsilon} \), respectively.

Under this model, the measured expression intensity \( y \) is a linear combination of a normal random variable and a lognormal random variable. The normal component \( (\alpha + \varepsilon) \) dominates at low expression levels, while the lognormal component \( (\mu e^{\eta}) \) dominates at high expression levels. The two-component model fits microarray data well at both low and high expression levels. Based on this model, we derive a generalized likelihood ratio (GLR) test to identify differentially expressed genes from microarray data.

**DATA**

\( y_{ijc} \) is the observed expression intensity of gene \( i \) in array \( j \) \((i = 1, \ldots, n; j = 1, \ldots, d_1)\) for the control sample. \( y_{ije} \) is the observed expression intensity of gene \( i \) in array \( j \) \((i = 1, \ldots, n; j = 1, \ldots, d_2)\) for the experimental sample.

The two-component model is

\[
y_{ijc} = \alpha_c + \mu_{ic} e^{\theta} + \varepsilon_{ijc} \quad \text{for the control sample, and} \]

\[
y_{ije} = \alpha_e + \mu_{ie} e^{\theta} + \varepsilon_{ije} \quad \text{for the experimental sample}
\]

where \( \eta_{ijc} \sim N(0, \sigma^2_{\eta c}), \varepsilon_{ijc} \sim N(0, \sigma^2_{\varepsilon c}), \eta_{ije} \sim N(0, \sigma^2_{\eta e}), \varepsilon_{ije} \sim N(0, \sigma^2_{\varepsilon e}). \)

To determine whether gene \( i \) is differentially expressed between control and experimental samples is equivalent to test the null hypothesis

\[
H_0: \mu_{ic} = \mu_{ie} \quad \text{against} \quad H_1: \mu_{ic} \neq \mu_{ie}.
\]

**The generalized likelihood ratio test**

The GLR test is a generalization of the Neyman–Pearson test. It provides a desirable test in many applications.

To simplify the analysis, we transform data:

\[
y'_{ijc} = y_{ijc} - \alpha_c = \mu_{ic} e^{\theta} + \varepsilon_{ijc} \quad \text{(2)}
\]

\[
y'_{ije} = y_{ije} - \alpha_e = \mu_{ie} e^{\theta} + \varepsilon_{ije} \quad \text{(3)}
\]

The density functions of \( y'_{ijc} \) and \( y'_{ije} \) can be derived by convolution formula

\[
f_{y'_{ijc}}(y) = \frac{1}{2\pi \sigma_{\varepsilon c} \sigma_{\eta c}} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ic})^2}{2\sigma_{\eta c}^2} - \frac{(y - e^{x})^2}{2\sigma_{\varepsilon c}^2} \right] dx
\]

(4)

\[
f_{y'_{ije}}(y) = \frac{1}{2\pi \sigma_{\varepsilon e} \sigma_{\eta e}} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ie})^2}{2\sigma_{\eta e}^2} - \frac{(y - e^{x})^2}{2\sigma_{\varepsilon e}^2} \right] dx
\]

(5)

The likelihood functions for control and experimental samples are defined by:

\[
L(\mu_{ic}; y'_{11c} \cdots y'_{d_1c}) = \prod_{j=1}^{d_1} \frac{1}{2\pi \sigma_{\varepsilon c} \sigma_{\eta c}} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ic})^2}{2\sigma_{\eta c}^2} - \frac{(y'_{ijc} - e^{x})^2}{2\sigma_{\varepsilon c}^2} \right] dx
\]

(6)

\[
L(\mu_{ie}; y'_{11e} \cdots y'_{d_2e}) = \prod_{j=1}^{d_2} \frac{1}{2\pi \sigma_{\varepsilon e} \sigma_{\eta e}} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ie})^2}{2\sigma_{\eta e}^2} - \frac{(y'_{ije} - e^{x})^2}{2\sigma_{\varepsilon e}^2} \right] dx
\]

(7)

The GLR test statistic for hypothesis \( H_0: \mu_{ic} = \mu_{ie} \) is defined by:

\[
\lambda_i = \max_{\mu_{ic} = \mu_{ie} > 0} \prod_{j=1}^{d_1} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ic})^2}{2\sigma_{\eta c}^2} - \frac{(y'_{ijc} - e^{x})^2}{2\sigma_{\varepsilon c}^2} \right] dx \times \prod_{j=1}^{d_2} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ie})^2}{2\sigma_{\eta e}^2} - \frac{(y'_{ije} - e^{x})^2}{2\sigma_{\varepsilon e}^2} \right] dx
\]

\[
\times \max_{\mu_{ic} > \mu_{ie} > 0} \prod_{j=1}^{d_1} \int_{-\infty}^{\infty} \exp \left[ -\frac{(x - \log \mu_{ie})^2}{2\sigma_{\eta e}^2} - \frac{(y'_{ije} - e^{x})^2}{2\sigma_{\varepsilon e}^2} \right] dx
\]

(8)

The parameters \( \alpha_c, \alpha_e, \sigma^2_{\eta c}, \sigma^2_{\eta e}, \sigma^2_{\varepsilon c}, \sigma^2_{\varepsilon e} \) are estimated using the procedure proposed by Rocke and Durbin (2001), \( \mu_{ic} \) and \( \mu_{ie} \) are estimated using maximum-likelihood estimation (MLE). We have implemented the calculation of \( \lambda_i \) in software called GLR.

**Distribution of GLR test statistic**

As a property of the GLR Test, \(-2 \log \lambda_i \) is approximately distributed as \( \chi^2(r) \) \((r = 1 \text{ in our case})\), when sample size is large. However, when sample size is small, which is the
and those of $\chi^2$ as small as one.

We generate 1000 samples of size $d_1 + d_2$, using model (1), with values of $\mu$ ranging from 0 to 0.01. The parameters used are $\alpha = 0$, $\sigma^2 = 0.0712$ and $\sigma^2 = 1.05 \times 10^{-10}$ for both control and experimental samples. The range of $\mu$ and the values of parameters are estimated from a data set containing bacterial *Escherichia coli* gene expression profile (Arfin et al., 2000). We compute the likelihood ratio test statistic for each sample and perform $\chi^2$ goodness-of-fit test and Cramer–Von Mises (CVM) goodness-of-fit test to test the hypothesis: $H_0: -2 \log \lambda \sim \chi^2(1)$. The results are shown in Table 1.

The results suggest that $-2 \log \lambda$ is approximately distributed as $\chi^2(1)$ at all expression levels and with sample sizes as small as one.

To determine if the distribution of $-2 \log \lambda$ and $\chi^2(1)$ are similar enough so we can choose the critical value of $-2 \log \lambda$ for hypothesis testing based on $\chi^2(1)$ distribution. We generate 10000 samples of size $d_1 + d_2$ with $\mu_i$ set at several different levels, determine the 99th and 95th percentiles of the distribution of $-2 \log \lambda$, and compare them with those of $\chi^2(1)$ distribution. The results are shown in Table 2.

The 99th and 95th percentiles of the distribution of $-2 \log \lambda$ and those of $\chi^2(1)$ distribution are very similar when $d \geq 2$ and $\mu_i \geq 10^{-4}$, while when $d = 1$ or $\mu_i \leq 10^{-5}$, the percentiles of the distribution of $-2 \log \lambda$ are much smaller than those of $\chi^2(1)$ distribution. Therefore, when the sample size is greater than or equal to 2 and the expression level is significantly larger than the base line error $\sigma_e$, it is fairly safe to assume that $-2 \log \lambda$ is distributed as $\chi^2(1)$ and determine the critical value for hypothesis tests accordingly. When the sample size is 1 or the expression level is not significantly larger than $\sigma_e$, using the critical value determined by $\chi^2(1)$ distribution, which is larger than the true critical value, makes the hypothesis test conservative, hence less powerful.

### Simulation study to compare the performance of the GLR test, the $t$-test and the fold-change method

The likelihood ratio test is the uniformly most powerful test for simple hypothesis $H_0: \mu = \mu_0$ versus $H_1: \mu = \mu_1$.

For composite hypothesis, its properties are not determined. We perform a simulation study to compare its performance with two widely used methods, the fold-change method and the two-sample $t$-test.

To compare the Type I error rates of three methods, we generated 10000 samples of size $d_1 + d_2$ using model (1) with $\mu_i = \mu_c = \mu_e$ set at several different levels. We compute the GLR test statistic $-2 \log \lambda$, the $t$-test statistics $t$ and the fold-change for each sample and calculate the ratio of samples whose statistics are in the critical region for each test, which is an estimator of Type I error rate of the tests. The size of the test is fixed at 0.05. The results are shown in Table 3.

The Type I error rates of the GLR test and the two-sample $t$-test estimated by simulation are very close to the sizes of the test, and they are not changed significantly when the sample size decreases from 4 to 2, except when $\mu_0 = 10^{-5}$, as we discussed before, when $\mu_0 = 10^{-5}$ the GLR test is conservative when using critical value determined by $\chi^2(1)$ distribution.

### Table 1. $P$-values of $\chi^2$ and CVM test for hypothesis: $H_0: -2 \log \lambda \sim \chi^2(1)$

<table>
<thead>
<tr>
<th>$\mu_i$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$P$-value $\chi^2$</th>
<th>CVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00001</td>
<td>4</td>
<td>4</td>
<td>0.350</td>
<td>0.51</td>
</tr>
<tr>
<td>0.0001</td>
<td>2</td>
<td>2</td>
<td>0.287</td>
<td>0.47</td>
</tr>
<tr>
<td>0.00001</td>
<td>1</td>
<td>1</td>
<td>0.251</td>
<td>0.53</td>
</tr>
<tr>
<td>0.0001</td>
<td>1</td>
<td>1</td>
<td>0.237</td>
<td>0.45</td>
</tr>
<tr>
<td>0.001</td>
<td>1</td>
<td>1</td>
<td>0.335</td>
<td>0.39</td>
</tr>
<tr>
<td>0.01</td>
<td>1</td>
<td>1</td>
<td>0.456</td>
<td>0.18</td>
</tr>
</tbody>
</table>

### Table 2. The 99th and 95th percentiles of the distribution of $-2 \log \lambda$ and $\chi^2(1)$

<table>
<thead>
<tr>
<th>$\mu_i$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>99th percentile $\chi^2$</th>
<th>95th percentile $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>4</td>
<td>4</td>
<td>6.63</td>
<td>6.648</td>
</tr>
<tr>
<td>0.01</td>
<td>2</td>
<td>2</td>
<td>6.63</td>
<td>6.403</td>
</tr>
<tr>
<td>0.01</td>
<td>1</td>
<td>1</td>
<td>6.63</td>
<td>5.487</td>
</tr>
<tr>
<td>0.001</td>
<td>4</td>
<td>4</td>
<td>6.63</td>
<td>6.370</td>
</tr>
<tr>
<td>0.001</td>
<td>2</td>
<td>2</td>
<td>6.63</td>
<td>6.467</td>
</tr>
<tr>
<td>0.001</td>
<td>1</td>
<td>1</td>
<td>6.63</td>
<td>5.677</td>
</tr>
<tr>
<td>0.0001</td>
<td>4</td>
<td>4</td>
<td>6.63</td>
<td>6.374</td>
</tr>
<tr>
<td>0.0001</td>
<td>2</td>
<td>2</td>
<td>6.63</td>
<td>6.317</td>
</tr>
<tr>
<td>0.0001</td>
<td>1</td>
<td>1</td>
<td>6.63</td>
<td>5.539</td>
</tr>
<tr>
<td>0.00001</td>
<td>4</td>
<td>4</td>
<td>6.63</td>
<td>6.338</td>
</tr>
<tr>
<td>0.00001</td>
<td>2</td>
<td>2</td>
<td>6.63</td>
<td>5.997</td>
</tr>
<tr>
<td>0.00001</td>
<td>1</td>
<td>1</td>
<td>6.63</td>
<td>5.417</td>
</tr>
</tbody>
</table>

### Table 3. Estimated Type I error rate for GLR test, $t$-test and fold-change method

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$\mu_0$</th>
<th>Type I error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>GLR test</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.01</td>
<td>0.052</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.001</td>
<td>0.049</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.0001</td>
<td>0.050</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.00001</td>
<td>0.051</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.01</td>
<td>0.049</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.001</td>
<td>0.053</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.0001</td>
<td>0.045</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.00001</td>
<td>0.035</td>
</tr>
</tbody>
</table>
We apply the GLR test to a data set comparing gene expression between IHF strains and IHF− strains of bacterial E.coli (Arfin et al., 2000). This data set contains gene expression data for 1973 genes and four replicates are performed for both IHF+ and IHF− strains. We take the genome-wide significance level at $\alpha = 0.1$, and use the Bonferroni method to adjust the significance level for individual gene, which is $\alpha^* = 0.1/(1973 \times 2) = 2.53 \times 10^{-5}$. The GLR test detects 88 genes, while the $t$-test detects 10 genes, in consistence with the results from the simulation study that the GLR test is more powerful than the $t$-test.

To estimate the percentage of such genes identified by chance [the false discovery rate (FDR)], we generate control data sets by creating two new groups each of which contains two IHF+ and two IHF− samples. Thirty-six such control data sets are analyzed, averagely 2.1 genes are identified from each control data set by the GLR test, giving a FDR of 2.3%. This number is higher than what the Bonferroni adjusted significant level predicted, suggesting some deviation of microarray data from the two-component model. When the significant level is increased to allow the $t$-test to identify 88 genes, it gives a FDR of 12.1%. Significance analysis of microarrays (SAM) is a modified $t$-test used to analyze microarray data. It tends to give lower FDRs than conventional methods (Tusher et al., 2001). When parameter is set to allow SAM to identify 88 genes, it gives a FDR of 3.8%. Therefore, GLR has the lowest FDR among the three methods.

Another method to evaluate a statistical method is to test if it can detect the same set of positive genes over independently repeated experiments. In the IHF data set, four replicates are performed for both IHF+ and IHF− strains. We use different subsets of data to test the consistency of statistical methods. We choose four subsets of data with size 3 replicates and 12 subsets of data with size 2 replicates (every subset is different by at least two replicates). We apply the GLR test, SAM and $t$-test to each subset of the data, identify the top 100 genes, and determine how many genes are commonly identified in these semi-independently repeated experiments. In the case of three by three comparisons, the GLR test averagely identifies 81% genes in common, SAM identifies 67% genes in common, while the $t$-test identifies 59% genes in common. In the case of two by two comparisons, the GLR test averagely identifies 76% genes in common, SAM identifies 51% genes in common, while the $t$-test identifies 34% genes in common. So the GLR test is more consistent than SAM and the $t$-test, especially when the sample size is small.

**DISCUSSION**

Microarray technology has the potential to provide fundamental insights into biological processes from regulation of gene expression to development of cancer. However, the statistical tools to analyze microarray data are generally unsatisfactory and become the bottleneck in the development of this new technology. One major difficulty of statistical analysis of microarray data is the large number of potential sources of random and systematic measurement error in microarray studies. It is unrealistic to build a model that reflects all sources of error.
and gives accurate estimation of all parameters of interest. Therefore, a good model should be complicated enough to capture the most essential features of the data but not too complicated so that the implementation is still practical. Currently, most statistical analysis of microarray data depends on the assumption that the data is normally distributed with variances not dependent on the mean of the data (Pan, 2002). More and more studies suggest that microarray data violate these assumptions rather dramatically. Several alternative models have been proposed for the measurement error in microarray data (Chen et al., 1997; Ideker, 2001; Rocke and Durbin, 2001). These models all reflect the observation that the variance of expression data of a gene increases with its mean. Among these models, the two-component model proposed by Rocke and Durbin fits the microarray data best at both high and low levels (Durbin, 2002).

Based on the two-component model, we develop a GLR test to identify differentially expressed genes from microarray data, and implement this approach in software called GLR. The GLR test statistic $-2 \log \lambda$ is approximately distributed as $\chi^2(1)$ when the sample size is greater or equal to two. Thus, the critical region of the GLR test can be determined by the $\chi^2(1)$ distribution without any unknown, experiment-specific parameter.

We demonstrate that the GLR test is more powerful than commonly used methods like the $t$-test and the fold-change method on simulated data. The GLR test with a sample size 2 can achieve a higher power than the $t$-test with a sample size 4 does. When applied to real data from microarray study, the GLR test also identifies more positive genes than the $t$-test. We also show that the GLR test has a lower FDR than both the $t$-test and SAM. The GLR test is also more consistent than the $t$-test and SAM, giving similar results when microarray experiments are repeated. Biologists can use GLR test to identify more truly differentially expressed genes with smaller samples, saving a lot of time and money.

ACKNOWLEDGEMENTS

We thank Lajos Horvath, Davar Koshnevisan and Robert Brooks for critical reading of the manuscript and fruitful discussion.

REFERENCES